ESTIMATION FOR DIRICHLET MIXED MODELS

BY

STEVE LEEDS and ALAN E. GELFAND

TECHNICAL REPORT NO. 428
MAY 19, 1990

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ESTIMATION FOR DIRICHLET MIXED MODELS

Steve Leeds and Alan E. Gelfand

ABSTRACT

Dirichlet mixed models find wide application. Estimation is usually achieved through the method of moments. Here we present an iterative hybrid algorithm for obtaining the maximum likelihood estimate employing both modified Newton-Raphson and E-M methods. This successful MLE algorithm enables calculation of a jackknife MLE. Simulation comparison of the three estimates is provided. The MLE substantially improves upon the moments estimator particularly with increasing dimension. The jackknife MLE in turn offers dramatic improvement over the MLE.

1. INTRODUCTION

Mixture distributions afford a flexible, rich class of models. A general definition of a mixture distribution which appears in Everitt and Hand (1981) goes as follows:

Let $g(x \mid \theta)$ be a d-dimensional probability density function with respect to some measure μ on a m-dimensional parameter vector θ and let $H(\theta)$ be a m-dimensional cumulative distribution function. Then

$$f(x) = \int_{\Theta} g(x \mid \theta)\, dH(\theta) \tag{1.1}$$

is called a mixture density. H is called the mixing distribution. If H is discrete and assigns positive probability to only a finite number of points $(\theta_i : i = 1, \ldots, c)$ then we have a finite mixture where

$$f(x) = \sum_{i=1}^{c} H(\theta_i) \cdot g(x \mid \theta_i) \tag{1.2}$$

Throughout the literature on mixture distributions the goal has been to estimate H assuming a parametric form g. Finite mixtures date back to Pearson (1894) who attempted to estimate the five parameters in a mixture of two normal distributions. Detailed discussions of mixtures can be found in Titterington, et al. (1985) and Everitt and Hand (1981).

Identifiability of the mixture model is a crucial issue. Teicher (1961,1963) was the first to give a definitive answer to this problem. By definition, a class D of mixtures is said to be identifiable if and only if for all $f(x) \in D$ the equality a.e. μ of the two representations:

$$\int_{\Theta} g(x \mid \theta)\, dH^{*}(\theta) = \int_{\Theta} g(x \mid \theta)\, dH(\theta) \tag{1.4}$$

implies that $H^{*}(\theta) \equiv H(\theta)$.

We will assume that H itself is from a parametric family indexed by α and that the goal is to estimate α based on observations, x, from

$$f(x \mid \alpha) = \int_{\Theta} g(x \mid \theta) \cdot dH(\theta \mid \alpha) \tag{1.5}$$

Discrete $f(x \mid \alpha)$ are more commonly referred to as compound distributions. Here we will consider X to be a vector of counts and θ (which will be a vector within the unit simplex) will

characterize the probabilities that a particular count within the X-vector will be incremented. Models for g include the multinomial distribution having density

$$g_\theta(x_1,x_2,\ldots,x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i}, \qquad \sum_{i=1}^{k} x_i = n, \quad \sum_{i=1}^{k} \theta_i = 1 \tag{1.6}$$

and the negative-multinomial distribution, having density

$$g_\theta(x_1,x_2,\ldots,x_{k-1}) = \frac{(n + \sum_{j=1}^{k-1} x_j - 1)!}{(n-1)! \prod_{j=1}^{k} x_j!} \prod_{j=1}^{k-1} \theta_j^{x_j} (1 - \sum_{j=1}^{k-1} \theta_j)^n, \quad \sum_{j=1}^{k-1} \theta_j < 1 \tag{1.7}$$

In both the multinomial and negative multinomial cases the random variable is defined by a particular stopping rule on the generalized Bernoulli trials. For the multinomial case the random variable is observed when n generalized Bernoulli trials are completed. For the negative multinomial case the random variable is observed when a predesignated k^{th} cell fills to size n. More broadly for a specified stopping rule on the generalized Bernoulli trials we say that the resulting random vector, $\underline{X}$, of observed counts for the k cells, follows a general occupancy distribution. Examples of other potentially interesting stopping rules are : (1) Sample until both $X_i \geq r_i$ and $X_j \geq r_j$, , (2) Sample until either $X_i = r_i$ or $X_j = r_j$. Generally, if $\underline{X}$ is an outcome in the sample space of a general occupancy model

$$g_\theta(\underline{x}) = c(\theta)\, h(x) \prod_{i=1}^{k} \theta_i^{x_i} \tag{1.8}$$

The natural conjugate choice for H in this context would be the Dirichlet distribution.

$$h_\alpha(\underline{\theta}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} [\prod_{j=1}^{k-1} \theta_j^{\alpha_j - 1}] (1 - \sum_{j=1}^{k-1} \theta_j)^{\alpha_k - 1}, \quad \alpha_i > 0,\ i = 1,\ldots,k \tag{1.9}$$

The assumption of the mixing distribution being conjugate buys simplicity of form. But in addition Dalal and Hall (1983) point out that arbitrary mixture distributions can be satisfactorily approximated by considering mixtures of natural conjugate distributions. In the present case we obtain the Dirichlet (or compound) - multinomial distribution

$$f_\alpha(x_1,\ldots,x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \frac{(\prod_{j=1}^{k-1} \Gamma(x_j + \alpha_j))\, \Gamma(n - \sum_{j=1}^{k-1} x_j + \alpha_k)}{\Gamma(n + \sum_{1=1}^{k} \alpha_i)} \tag{1.10}$$

$$= \frac{n!}{\prod_{i=1}^{k} x_i!} \frac{\prod_{i=1}^{k-1} \prod_{r=0}^{x_i - 1} (\alpha_i + r) \prod_{r=0}^{n - \sum_{j=1}^{k-1} x_j - 1} (\alpha_k + r)}{\prod_{r=0}^{n-1} (\alpha_1 + \alpha_2 + \cdots + \alpha_k + r)}$$

and the Dirichlet (or compound) - negative Multinomial Distribution

$$f_{\alpha}(x_1, \ldots, x_{k-1}) = \frac{(n + \sum_{j=1}^{k-1} x_j - 1)!}{(n-1)! \prod_{j=1}^{k-1} x_j!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \frac{(\prod_{j=1}^{k-1} \Gamma(x_j + \alpha_j)) \Gamma(n + \alpha_k)}{\Gamma(n + \sum_{j=1}^{k-1} x_j + \sum_{i=1}^{k} \alpha_i)} \tag{1.11}$$

$$= \frac{(n + \sum_{j=1}^{k-1} x_j - 1)!}{(n-1)! \prod_{j=1}^{k-1} x_j!} \frac{\prod_{i=1}^{k-1} \prod_{r=0}^{x_i - 1} (\alpha_i + r) \prod_{r=0}^{n-1} (\alpha_k + r)}{\prod_{r=0}^{n + \sum_{j=1}^{k-1} x_j - 1} (\alpha_1 + \alpha_2 + \cdots + \alpha_k + r)}$$

Extension to a Dirichlet general occupancy distribution is apparent.

With regard to identifiability let Ω be the collection of all possible distinct events in the sample space and let $N(\Omega)$ be the cardinality of Ω . For any k dimensional α-vector we can construct $N(\Omega)$ equations which describe the probabilities for all simple events in Ω. Since all of the probabilities must sum to 1 we will have $N(\Omega) - 1$ independent equations and k unknowns. If we have at least as many equations as we have unknowns, we cannot choose any other α-vector that will generate the same probabilities. Therefore, the condition for an identifiable Dirichlet mixture will be

$$(N(\Omega) - 1) - k \geq 0 \tag{1.12}$$

Hence, the Dirichlet negative multinomial is always identifiable, while the Dirichlet multinomial is, provided $n > 1$.

Applications of these two mixture models are extensive in the literature. In particular see Leckenby and Kishi (1984), Rust and Leone (1984), Kalwani (1980), and Mosimann (1962,1963).

2. ESTIMATION APPROACHES

Method of moments estimation is most commonly used in the Dirichlet mixed models (see e.g. Mosimann (1962,1963), Johnson and Kotz (1969)). Maximum likelihood estimation requires a difficult numerical maximization and has been studied primarily in the Beta mixed (k = 2) case. (See e.g. Griffiths (1973), Smith (1983), Williams (1975)). Computation of MLE's is the primary issue of section 3. Effective computation of the MLE enables us to propose a jackknifed MLE as a third choice. Mean square error behavior of the jackknife estimator is extremely promising.

We first review the method of moments estimator. First and second moments associated with the Dirichlet multinomial in (1.9) are

$$EX_i = n \frac{\alpha_i}{\Sigma\alpha} \tag{2.1}$$

$$VarX_i = \frac{n + \Sigma\alpha}{1 + \Sigma\alpha} \left(n \frac{\alpha_i}{\Sigma\alpha} \left(1 - \frac{\alpha_i}{\Sigma\alpha}\right) \right) \tag{2.2}$$

$$Cov(X_i, X_{i'}) = \frac{n + \Sigma\alpha}{1 + \Sigma\alpha} \left(n \frac{\alpha_i}{\Sigma\alpha} \frac{\alpha_{i'}}{\Sigma\alpha} \right) \tag{2.3}$$

To obtain the moments estimates we would have the relationship $n\alpha_i / \Sigma\alpha = \bar{x}_i$. Since there is no constraint on the α_i's , we need an extra equation There is no unique criteria to determine this extra equation and therefore can be considered an ad hoc choice.

Mosimann (1962) noticed that the covariance structure for the Dirichlet-multinomial is just a constant times the covariance matrix for a multinomial with parameters $\left(\frac{\alpha_1}{\Sigma\alpha}, \ldots, \frac{\alpha_k}{\Sigma\alpha}\right)$.

$$\Sigma_{DM} = c \cdot \Sigma_{M(p_1', \ldots, p_k')} \tag{2.4}$$

$$\text{where } p_i' = \frac{\alpha_i}{\Sigma\alpha}, \quad c = \frac{n + \Sigma\alpha}{1 + \Sigma\alpha}$$

Using this relationship the generalized variance would be $|\Sigma_{DM}| = c^{k-1} \cdot |\Sigma_M|$. In order to avoid singular matrices, (k-1) terms of the covariance matrix are used.

For a sample $(X_{t1}, \ldots, X_{tk})$ t = 1,...,m (i.e. m replications) from this distribution, Mosimann showed that the MME's are of the form

$$\hat{\alpha}_i = \frac{\bar{x}_i(n - \hat{c})}{n(\hat{c} - 1)} \qquad i = 1, \ldots, k \tag{2.5}$$

$$\hat{c}^{k-1} = \frac{|\hat{\Sigma}_{DM}|}{|\Sigma_M|} = \frac{|S|}{\left(\prod_{j=1}^{k-1} \bar{x}_j\right)\left(n - \sum_{j=1}^{k-1} \bar{x}_j\right)^{n}} \tag{2.6}$$

Here |S| is the determinant of the sample covariance matrix

$$S_{ii} = S_i^2 = \sum_{t=1}^{m} \frac{(x_{ti} - \bar{x}_i)^2}{m} \qquad i = 1, \ldots, k-1 \tag{2.7}$$

$$S_{ij} = \sum_{t=1}^{m} \frac{(x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)}{m} \qquad i, j = 1, \ldots, k-1 \tag{2.8}$$

In order for α to be feasible, $1 < c < n$. If this condition is not satisfied then we will say that the MME does not exist for this mixture case.

First and second moments associated with the Dirichlet negative multinomial in (1.11) are

$$EX_j = n\frac{\alpha_j}{\alpha_k - 1}, \qquad \alpha_k > 1 \tag{2.9}$$

$$VarX_j = \frac{n + \alpha_k - 1}{\alpha_k - 2}\left(n\frac{\alpha_j}{\alpha_k - 1}\frac{\alpha_j + \alpha_k - 1}{\alpha_k - 1}\right), \qquad \alpha_k > 2 \tag{2.10}$$

$$Cov(X_j, X_{j'}) = \frac{n + \alpha_k - 1}{\alpha_k - 2}\left(n\frac{\alpha_j}{\alpha_k - 1}\frac{\alpha_{j'}}{\alpha_k - 1}\right), \qquad \alpha_k > 2 \tag{2.11}$$

$$j, j' = 1, \ldots, k-1$$

Again Mosimann (1963) exploited the covariance structure of the negative multinomial and Dirichlet-negative multinomial by observing that

$$\Sigma_{DNM} = c \cdot \Sigma_{NM(p_1', \ldots p_k')}$$

$$\text{where } p'_j = \alpha_j, \quad j = 1, \ldots, k-1, \qquad p'_k = \alpha_k - 1$$

to yield

$$\hat{\alpha}_i = \frac{\bar{x}_j}{n}\frac{(n + \hat{c})}{\hat{c} - 1} \qquad j = 1, \ldots, k-1 \tag{2.12}$$

$$\hat{\alpha}_k = \frac{(2\hat{c} + n - 1)}{\hat{c} - 1} \tag{2.13}$$

For some reason Mosimann did not present the generalized variance ratio estimate of c for the DNM case. Instead he chose c by ignoring the overall covariance structure and equated

$$TR(\Sigma_{DNM}) = c \cdot TR(\Sigma_{NM}) \tag{2.14}$$

An alternative estimate consistent with the Dirichlet Multinomial approach is one that accounts for the full covariance structure. Therefore we use

$$\hat{c}^{k-1} = \frac{|\hat{\Sigma}_{DNM}|}{|\hat{\Sigma}_{NM}|} = \frac{|S|}{(\prod_{j=1}^{k-1} \bar{x}_j)\ (n + \sum_{j=1}^{k-1} \bar{x}_j)/n} \tag{2.15}$$

We should remember that this choice of estimate is ad hoc. However, we will call the resulting estimate the MME, even though it is not uniquely determined. Again, if $c \leq 1$ we will say that the MME does not exist.

There has been no study up to this time which evaluates the performance of these estimates in both the DM and DNM cases. Since the estimates in both cases are rational functions of consistent estimates (DNM case $\alpha_k > 2$) , they too are consistent (Slutsky's theorem). Results of simulation studies will be presented in section 4.

With regard to maximum likelihood estimation for multinomial and negative multinomial distributions, MLE's and MME's are the same. In the mixture case, MLE's cannot be written out in closed form. In order to proceed with maximum likelihood estimation it is sometimes more convenient to write the compound distribution of interest under a particular reparametrization. The reasoning behind this is that we will ultimately need an iterative procedure to obtain the MLE. If we can choose a parametrization such that the parameter estimates do not 'vary much in the region of best-fitting models', then we will have a more efficient iterative procedure. These new parameters are called stable parameters. Ross (1970) discusses maximum likelihood in this context.

In the Dirichlet multinomial model we reparametrize to

$$f_{\underline{\mu},\,\theta}(\underline{x}) = \binom{n}{x_1, \ldots, x_k} \frac{\prod_{i=1}^{k-1} \prod_{r=0}^{x_i - 1} (\mu_i + r\theta) \prod_{r=0}^{n - \sum_{j=1}^{k-1} x_j - 1} (1 - \sum_{j=1}^{k-1} \mu + r\theta)}{\prod_{r=0}^{n-1} (1 + r\theta)}$$

$$0 < \mu_i < 1, \quad i = 1, \ldots, k-1, \quad \theta > 0 \tag{2.16}$$

$$\text{where } \mu_i = \frac{\alpha_i}{\Sigma\alpha} \qquad i = 1, \ldots, k-1 \text{ and } \quad \theta = \frac{1}{\Sigma\alpha} \tag{2.17}$$

Under this parametrization μ_i can be thought of as the mean parameter of the original p_i , and θ can be thought of as a shape parameter. Griffiths (1973) seems to have been the first to use this representation for the (k = 2) case. Under this parametrization , $f_{\underline{u},\,\theta}(\underline{x})$ is exactly a multinomial with parameters $(\mu_1, \mu_2, \ldots, 1 - \sum^{k-1} \mu_j)$. Thus departures from $\theta = 0$ suggest departures from pure multinomial variation. θ in this setting is sometimes called an overdispersion parameter. From expression (2.4) we notice that

$$c = \frac{n\theta + 1}{\theta + 1} \tag{2.18}$$

so when $\theta = 0$, $c = 1$.

Skellam (1948) examined the log of expression (2.26) for the case of ($k = 2$) and took the natural 'derivative log likelihood' approach. He proposed a recursive procedure through the derivative log gamma or digamma funtion. Since the digamma function must be approximated, this procedure for the general case is not appealing.

Using the reparametrization, (2.17), the Dirichlet negative multinomial becomes

$$f_{\underline{\mu},\,\theta}(\underline{x}) = \binom{n + \sum_{j=1}^{k-1} x_j - 1}{n-1} \frac{\prod_{i=1}^{k-1} \prod_{r=0}^{x_i - 1} (\mu_i + r\theta) \prod_{r=0}^{n-1} (1 - \sum_{j=1}^{k-1} \mu + r\theta)}{\prod_{r=0}^{n + \sum_{j=1}^{k-1} x_j - 1} (1 + r\theta)}$$

$$0 < \mu_i < 1, \quad i = 1, \ldots, k-1, \quad \theta > 0 \tag{2.19}$$

When $\theta = 0$ we have a pure negative multinomial distribution, so θ again conveys departures from negative multinomial variation. In this case, due to the constraint on the second moments (i.e. $\alpha_k > 2$), $0 < \theta < 0.5$ and

$$c = \frac{1 - \sum_{j=1}^{k-1} \mu_j + (n-1)\theta}{1 - \sum_{j=1}^{k-1} \mu_j - 2\theta}$$

The likelihood system of equations in the DM ($\mu_1, \ldots, \mu_{k-1}, \theta$) case becomes

$$\frac{\partial \mathscr{L}}{\partial \mu_i} = \sum_{t=1}^{m} \sum_{r=0}^{x_{ti}-1} \frac{1}{\mu_i + r\theta} - \sum_{t=1}^{m} \sum_{r=0}^{x_{tk}-1} \frac{1}{1 - \sum_{j=1}^{k-1} \mu_j + r\theta} = 0 \qquad i = 1, \ldots, k-1$$

$$\frac{\partial \mathscr{L}}{\partial \theta} = \sum_{t=1}^{m} \sum_{i=1}^{k-1} \sum_{r=0}^{x_{ti}-1} \frac{r}{\mu_i + r\theta} - \sum_{t=1}^{m} \sum_{r=0}^{x_{tk}-1} \frac{r}{1 - \sum_{j=1}^{k-1} \mu_j + r\theta}$$

$$- \sum_{t=1}^{m} \sum_{r=0}^{n_t - 1} \frac{r}{1 + r\theta} = 0 \tag{2.20}$$

The likelihood system of equations in the DNM ($\mu_1, \ldots, \mu_{k-1}, \theta$) case becomes

$$\frac{\partial \mathscr{L}}{\partial \mu_i} = \sum_{t=1}^{m} \sum_{r=0}^{x_{ti}-1} \frac{1}{\mu_i + r\theta} - \sum_{t=1}^{m} \sum_{r=0}^{n_t - 1} \frac{1}{1 - \sum_{j=1}^{k-1} \mu_j + r\theta} = 0 \qquad i = 1, \ldots, k-1$$

$$\frac{\partial \mathscr{L}}{\partial \theta} = \sum_{t=1}^{m} \sum_{i=1}^{k-1} \sum_{r=0}^{x_{ti}-1} \frac{r}{\mu_i + r\theta} - \sum_{t=1}^{m} \sum_{r=0}^{n_t - 1} \frac{r}{1 - \sum_{j=1}^{k-1} \mu_j + r\theta}$$

$$- \sum_{t=1}^{m} \sum_{r=0}^{n_t + \sum_{j=1}^{k-1} x_{tj} - 1} \frac{r}{1 + r\theta} = 0 \tag{2.21}$$

It is easily seen that for both (2.20) and (2.21) the likelihood equations do not yield a solution in closed form ; an iterative procedure is required. Once MLE's for ($\underline{\mu}$, θ) are obtained, we can easily

convert them to MLE's for α_i. For both distributional models we can obtain straightforward expressions for Fisher's information matrix by taking a second derivative of the likelihood. See Leeds (1987) for details. It is important to note that unequal n_i can be used in maximum likelihood estimation, but it is unclear how to proceed for moments estimation. Ad hoc weighting procedures have been suggested for the (k = 2) case (see e.g. Kleinman (1970)). To allow comparisons we will take n_i equal in our simulation studies.

We also propose jackknifing of the MLE. The jackknife idea dates back to Quenouille (1956) and is thoroughly discussed in Efron (1982). We consider the jackknife procedure for two purposes. First, we hope to obtain a bias reduction, hence, a possibly better MSE performing estimator (see Schucany, et al. 1971). Second, we wish to study the performance of the jackknife estimate for possible confidence interval development.

We recall that given a sample of size m and a point estimate $\hat{\Phi}$ of unknown parameter Φ, the jackknife constructs what are known as "pseudovalues" which are defined by the relationship

$$\Phi_i^* = m\hat{\Phi}_{ALL} - (m-1)\hat{\Phi}_{(i)} \qquad i = 1,2,\ldots,m \tag{2.22}$$

where $\hat{\Phi}_{ALL}$ is the original estimate with all observations included and $\hat{\Phi}_{(i)}$ is the computed value of the estimate with the i^{th} observation removed. We would then compute the average and standard deviation for the set of pseudovalues and call them Φ^* and s_* respectively. Here

$$\Phi^* = \frac{1}{m}\sum_{i=1}^{m}\Phi_i^* \tag{2.23}$$

$$s_*^2 = \frac{\sum_{i=1}^{m}(\Phi_i^* - \Phi^*)^2}{m-1} \tag{2.24}$$

Our jackknife variance estimate of $Var(\hat{\Phi}_{ALL})$ (possibly of $Var(\Phi^*)$) would be $s_J^2 = m^{-1}s_*^2$ and we would then construct the confidence set

$$\hat{\Phi}_{ALL} \pm t_{m-1} \cdot s_J \tag{2.25}$$

Successful jackknifing depends upon successful computation of pseudovalues emphasizing the need for an effective iterative MLE procedure. We develop such a procedure in the next section.

3. ITERATIVE PROCEDURES FOR MLE'S

In this section we investigate in detail computational methods for obtaining the MLE. When maximizing the log likelihood function it was proposed that we solve the likelihood system.

$$\frac{\hat{c}\mathcal{L}}{\hat{c}\mu_i} = 0 \qquad i = 1, \ldots, k-1$$

$$\frac{\hat{c}\mathcal{L}}{\hat{c}\theta} = 0$$

Recalling that the solution to these equations (if it exists) can not be represented in closed form for the mixture cases discussed, numerical methods sometimes referred to as "root finders" must be considered. The most common methods for determining the roots for a system of nonlinear equations are the Newton-Raphson Method (NR) and the Method of Scoring (MS).

If we let $\Phi = (\Phi_1, \ldots, \Phi_k)$ denote the vector parameter and $\mathcal{L}(\Phi)$ be the log likelihood evaluated at Φ , then the NR algorithm at iteration (r + 1) is defined by

$$\Phi^{(r+1)} = \Phi^{(r)} - \gamma_r [D^2 \mathcal{L}(\Phi^{(r)})]^{-1} D\mathcal{L}(\Phi^{(r)}), \qquad r = 0,1,2, \ldots \tag{3.1}$$

and the MS algorithm is defined by

$$\Phi^{(r+1)} = \Phi^{(r)} - \gamma_r [I(\Phi^{(r)})]^{-1} D\mathcal{L}(\Phi^{(r)}), \qquad r = 0,1,2, \ldots \tag{3.2}$$

In both cases the non negative constant γ_r can be thought of as a damping term ($\gamma_r = 1$ is the usual version). D and D^2 are differential operators representing generalized first and second derivatives respectively, and $I(\Phi^{(r)})$ is Fisher's information matrix. We consider only the NR algorithm since $D^2\mathcal{L}(\Phi^{(r)})$ is available in closed form. The MS algorithm requires the numerical computation of an expected matrix at each stage. In fact, we use a modified Newton Raphson approach to avoid the required matrix inversion in (3.1) at each iteration. There are many different versions of the modified Newton-Raphson method with the least attention given to the simplest version. This version can be constructed by computing only the second derivatives on each coordinate separately and setting the mixed partials to zero.

Let

$$H_{ii} = D_{ii}^2 \mathcal{L}(\Phi) \qquad i = 1,2, \ldots, k \tag{3.3}$$

$$H_{ij} = 0, \qquad i \neq j \quad = 1, \ldots, k$$

Now

$$H_{ii}^{-1} = (D_{ii}^2 \mathcal{L}(\Phi))^{-1} = \frac{1}{\frac{\hat{c}^2}{\hat{c}\Phi_i^2}\mathcal{L}(\Phi)} \qquad i = 1,2, \ldots, k \tag{3.4}$$

$$H_{ij}^{-1} = 0, \qquad i \neq j \quad = 1, \ldots, k$$

yielding the system of equations

$$\Phi_1^{(r+1)} = \Phi_1^{(r+1)} - \left. \frac{\frac{\hat{c}}{\hat{c}\Phi_1}\mathcal{L}(\Phi)}{\frac{\hat{c}^2}{\hat{c}\Phi_1^2}\mathcal{L}(\Phi)} \right|_{\Phi = \Phi^{(r)}}$$

$$\vdots$$

$$\Phi_k^{(r+1)} = \Phi_k^{(r+1)} - \frac{\frac{\hat{c}}{\hat{c}\Phi_k}\mathscr{L}(\Phi)}{\frac{\hat{c}^2}{\hat{c}\Phi_k^2}\mathscr{L}(\Phi)}\Bigg|_{\Phi=\Phi^{(r)}} \tag{3.5}$$

which can be thought of as k versions of the univariate Newton-Raphson method. This diagonal version, where equations would be updated as each new coordinate becomes available, will be referred to as the Modified Newton Raphson (MNR) algorithm. We remark that (1) Convergence of the NR or MNR algorithm is typically dependent on the starting solution $\Phi^{(0)}$. (2) When MNR converges, it converges quickly. (3) The MNR algorithm may not converge to the root Φ^* even when started 'close' to Φ^*. (4) For what follows we take the parametrization $\Phi = (\mu_1, \mu_2, \ldots, \mu_{k-1}, 0)$ in which case the MNR algorithm becomes

$$\mu_i^{(r+1)} = \mu_i^{(r)} - \frac{\frac{\hat{c}}{\hat{c}\mu_i}\mathscr{L}(\Phi)}{\frac{\hat{c}^2}{\hat{c}\mu_i^2}\mathscr{L}(\Phi)}\Bigg|_{\Phi=\Phi^{(r)}} \qquad i = 1,2,\ldots,k-1 \tag{3.6}$$

$$\theta^{(r+1)} = \theta^{(r)} - \frac{\frac{\hat{c}}{\hat{c}\theta}\mathscr{L}(\Phi)}{\frac{\hat{c}^2}{\hat{c}\theta^2}\mathscr{L}(\Phi)}\Bigg|_{\Phi=\Phi^{(r)}} \tag{3.7}$$

For the Dirichlet multinomial model, first partials are given in (2.19) with

$$\begin{aligned}
\frac{\hat{c}^2}{\hat{c}\mu_i^2}\mathscr{L}(\Phi) &= -\sum_{t=1}^{m}\sum_{r=0}^{x_{ti}-1}\frac{1}{(\mu_i + r\theta)^2} - \sum_{t=1}^{m}\sum_{r=0}^{x_{tk}-1}\frac{1}{(1-\sum_{j=1}^{k-1}\mu_j + r\theta)^2} \\
\frac{\hat{c}^2}{\hat{c}\theta^2}\mathscr{L}(\Phi) &= -\sum_{t=1}^{m}\sum_{i=1}^{k-1}\sum_{r=0}^{x_{ti}-1}\frac{r^2}{(\mu_i + r\theta)^2} - \sum_{t=1}^{m}\sum_{r=0}^{x_{tk}-1}\frac{r^2}{(1-\sum_{j=1}^{k-1}\mu_j + r\theta)^2} \\
&\quad + \sum_{t=1}^{m}\sum_{r=0}^{n_t-1}\frac{r^2}{(1+r\theta)^2}
\end{aligned} \tag{3.8}$$

For the Dirichlet negative multinomial model, first partials are given in (2.20) with

$$\begin{aligned}
\frac{\hat{c}^2}{\hat{c}\mu_i^2}\mathscr{L}(\Phi) &= -\sum_{t=1}^{m}\sum_{r=0}^{x_{ti}-1}\frac{1}{(\mu_i + r\theta)^2} - \sum_{t=1}^{m}\sum_{r=0}^{n_t-1}\frac{1}{(1-\sum_{j=1}^{k-1}\mu_j + r\theta)^2} \\
\frac{\hat{c}^2}{\hat{c}\theta^2}\mathscr{L}(\Phi) &= -\sum_{t=1}^{m}\sum_{i=1}^{k-1}\sum_{r=0}^{x_{ti}-1}\frac{r^2}{(\mu_i + r\theta)^2} - \sum_{t=1}^{m}\sum_{r=0}^{n_t-1}\frac{r^2}{(1-\sum_{j=1}^{k-1}\mu_j + r\theta)^2} \\
&\quad + \sum_{t=1}^{m}\sum_{r=0}^{n_t+\sum_{j=1}^{k-1}x_{tj}-1}\frac{r^2}{(1+r\theta)^2}
\end{aligned} \tag{3.9}$$

A major problem with the NR or MNR algorithm is the need for a good starting solution. The MME would seem to be a reasonable starting solution, but as mentioned before, there are many cases where we cannot compute a moments estimate. Even if we can compute the MME, what recourse do we have if this starting solution causes the algorithm to diverge?

The EM algorithm (Dempster, et al., (1977)) offers an alternative approach The 'algorithm' was originally proposed for the treatment of incomplete data, but can accommodate many other

situations. The EM algorithm generates from some starting solution $\Phi^{(0)}$ a sequence $\{\Phi^{(r)}\}$ of estimates in the following steps

$$\text{E-STEP:} \quad \text{Evaluate } E(\mathcal{L}(\Phi) | y, \Phi^{(r)}) = Q(\Phi, \Phi^{(r)}) \tag{3.10}$$

$$\text{M-STEP:} \quad \text{Find } \Phi = \Phi^{(r+1)} \text{ to maximize } Q(\Phi, \Phi^{(r)}) \tag{3.11}$$

One of the more appealing properties of the EM algorithm is that under mild conditions (see e.g. Wu, 1983) each succesive iterate increases the likelihood function.

When the complete data likelihood comes from a regular exponential family represented by

$$f(y|\Phi) = b(y) \exp\{\Phi \cdot t(y)^T\} / a(\Phi) \tag{3.13}$$

where $t(y)^T$ denotes a (k x 1) 'complete data' sufficient statistic, the E and M steps take on a more explicit form.

E-STEP : Estimate the complete data sufficient statistic by finding

$$t^{(r)} = E(t(y) | x, \Phi^{(r)}) \tag{3.14}$$

M-STEP : Determine $\Phi^{(r+1)}$ as the solution of the equations

$$E(t(y) | \Phi) = t^{(r)} \tag{3.15}$$

The M-STEP is a maximization step because this condition must hold when obtaining a maximum under a regular exponential family model. Cycling back and forth between E and M steps should yield the MLE when and if the iterative sequence stabilizes. At the MLE Φ^*, the following relationship holds:

$$E(t(y) | x, \Phi^*) = E(t(y) | \Phi^*) \tag{3.16}$$

This equality of conditional to unconditional expectation at the MLE has also been noticed by other authors (Baum, et al., 1970, Orchard and Woodbury, 1972, Sundberg, 1974).

Under Dirichlet mixture we have the 'incomplete data' $X = \{(x_1, \ldots, x_k)\}$ or $\{(x_1, \ldots, x_{k-1})\}$ as the case may be, and the 'complete data' $Y = \{(x_1, \ldots, x_k, p_1, \ldots, p_k)\}$. Thus

$$f(y|\Phi) = f(\underline{x}, \underline{p} | \Phi) = f(\underline{x} | \underline{p}) \cdot D(\underline{p} | \underline{\alpha}) \tag{3.17}$$

where $f(\underline{x} | \underline{p})$ represents the multinomial or negative multinomial and $D(\underline{p} | \underline{\alpha})$ represents the Dirichlet distribution. Hence the distribution $f(y | \Phi)$ is an exponential family distribution. If we now sample $(x_{t1}, \ldots, x_{tk})$, $t = 1,2, \ldots, m$ we have

$$L(\Phi) = \prod_{t=1}^{m} [f(x_{t1}, \ldots, x_{tk} | p_{t1}, \ldots, p_{tk}) \cdot D(p_1, \ldots, p_k | \alpha_1, \ldots, \alpha_k)] \tag{3.18}$$

$$= [\prod_{t=1}^{m} f(\underline{x}_t, \underline{p}_t)] \cdot [\prod_{t=1}^{m} \prod_{i=1}^{k} p_{ti}^{\alpha_i - 1}] \cdot \exp\{\sum_{i=1}^{k} \alpha_i \sum_{t=1}^{m} \log p_{ti}\} / \left(\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}\right)^{-m}$$

where $\Phi = (\alpha_1, \ldots, \alpha_k)$ and $t(y) = t(p) = \sum_{t=1}^{m} \log p_{ti}$ $\quad i = 1, \ldots, k.$

It is interesting to note that $t(p)$ is always the same under Dirichlet mixture and thus does not depend on the original distribution being mixed.

To put together the E and M steps we must compute

$$E(t(p) | \Phi) = E(\sum_{t=1}^{m} \log p_{ti} | \underline{\alpha}) \quad \textit{and} \quad E(t(p) | x, \Phi) = E(\sum_{t=1}^{m} \log p_{ti} | \underline{x}_t, \underline{\alpha}) \tag{3.19}$$

Now $E(\sum_{i=1}^{m} \log p_{ti} \mid \underline{\alpha}) = \sum_{i=1}^{m} E(\log p_{ti} \mid \underline{\alpha}) = m\, E(\log p_{1i} \mid \underline{\alpha})$

Using the expansion $\log p_{1i} = -\sum_{j=1}^{\infty} \frac{(1-p_{1i})^j}{j}$ after some manipulation yields

$$E(\sum_{i=1}^{m} \log p_{ti} \mid \underline{\alpha}) = -m \cdot \sum_{j=1}^{\infty} \frac{1}{j} \cdot \prod_{s=1}^{j} \left[\frac{\Sigma\alpha - \alpha_i + s - 1}{\Sigma\alpha + s - 1}\right] \tag{3.20}$$

In order to evaluate $E(\sum_{i=1}^{m} \log p_{ti} \mid \underline{x}_t, \underline{\alpha})$ we notice that the distribution of $\underline{p} \mid \underline{x}$ is also Dirichlet. Similar calculations to that yielding (3.20) produces

$$E(\sum_{i=1}^{m} \log p_{ti} \mid \underline{x}_t, \underline{\alpha}) = -\sum_{i=1}^{m} \sum_{j=1}^{\infty} \frac{1}{j} \prod_{s=1}^{j} \left[\frac{\Sigma\gamma(t) - \gamma_i(t) + s - 1}{\Sigma\gamma(t) + s - 1}\right] \tag{3.21}$$

where $\gamma_i(t) = X_{ti} + \alpha_i, \quad i = 1, \ldots, k$

Observing (3.15) and (3.20) we see that in our case the M-step does not admit a close form solution. The M-step would have to be solved iteratively within each iteration. Instead we use the necessary condition (3.16). Solving the stable point problem in (3.16) will be called the modified EM algorithm (or MEM algorithm). Solving the EM algorithm will solve (3.16). However, solving (3.16) does not necessarily provide the solution generated by the EM algorithm, unless the solution to (3.16) is unique. In general under Dirichlet mixture the MEM equations (3.16) are

$$\sum_{i=1}^{m} \sum_{j=1}^{\infty} \left[E(1-p_{ti})^j \mid \underline{x}_t, \underline{\alpha} \; - E(1-p_{ti})^j \mid \underline{\alpha} \; \right] = 0 \qquad i = 1, \ldots, k$$

When using the EM algorithm to obtain an exact solution for the MLE it is well known that convergence to a solution is extremely slow. However, Redner and Walker (1984) point out that a quick climbing of the likelihood surface usually occurs in only a few iterations. We hope to retain this feature with our proposed MEM algorithm.

By a quick inspection we can see that $\theta = 0$ is a solution to the MEM equations in both cases. However, this result is never achieved unless $\theta = 0$ is used as a starting solution. Explicit solutions are not available and therefore a root finding method such as the MNR method can be used here. The infinite summations can be truncated to obtain approximate solutions.

To solve the MEM equations we would use the (μ, θ) parametrization and construct the MNR system in the same way it was constructed for the orginal likelihood equations. In this case $\frac{\partial}{\partial \Phi} \mathcal{L}(\Phi)$ is replaced by expression (3.16) yielding the system

$$\mu_i^{(r+1)} = \mu_i^{(r)} - \left. \frac{E(t(y) \mid x, \Phi) - E(t(y) \mid \Phi)}{\frac{\partial}{\partial \Phi}\left[E(t(y) \mid x, \Phi) - E(t(y) \mid \Phi)\right]} \right|_{\Phi = \Phi^{(r)}} \qquad i = 1, \ldots, k-1$$

$$\theta^{(r+1)} = \theta^{(r)} - \left. \frac{E(t(y) \mid x, \Phi) - E(t(y) \mid \Phi)}{\frac{\partial}{\partial \Phi}\left[E(t(y) \mid x, \Phi) - E(t(y) \mid \Phi)\right]} \right|_{\Phi = \Phi^{(r)}} \tag{3.22}$$

To compute derivatives we use the fact that

$$\frac{\partial}{\partial \Phi} \prod_{s=1}^{j} h_s(\Phi) = \frac{\partial}{\partial \Phi} \exp\{\sum_{s=1}^{j} \log h_s(\Phi)\}$$

$$= \exp\{\sum_{s=1}^{j} \log h_s(\Phi)\} \cdot [\sum_{s=1}^{j} \frac{\frac{\partial}{\partial \Phi} h_s(\Phi)}{h_s(\Phi)}] = [\prod_{s=1}^{j} h_s(\Phi)] \cdot [\sum_{s=1}^{j} \frac{\frac{\partial}{\partial \Phi} h_s(\Phi)}{h_s(\Phi)}]$$

We can substitute the appropriate expressions for the derivative in the MNR system because the MEM equations are just functions of products of this type. For efficient computation of all terms, recurrence relationships for the product functions and derivatives can be created for increasing j.

Finally we state our MLE algorithm. It can be thought of as a hybrid algorithm given by the following steps:

1) Choose a starting solution $\Phi^{(0)}$. If Φ_{MME} exists then $\Phi^{(0)} = \Phi_{MME}$. If $\Phi^{(0)}$ does not exist then starting solution $(\varepsilon, \varepsilon, \ldots, \varepsilon, \frac{1}{k})$ is used.

2) Iterate using the MNR equations in (3.7) with derivatives given by expression (3.8) or (3.9).

3a) If step 2 yields a converging sequence $\{\Phi_{(r)}\} \rightarrow \Phi^*$ then $\Phi_{MLE} = \Phi^*$.

3b) If step 2 diverges then we run about 20 MEM iterations starting at the last iterate generated by the previous MEM run. If MEM is being run for the first time then we can start at $\Phi^{(0)}$. MEM is intended to point failed MNR starting solutions in the right direction. Return to step 2 after 20 MEM iterations are completed.

In concluding this section we address the question of whether the proposed MLE algorithm obtains the MLE. To do so we investigate the likelihood surface and ask the following questions:

A) If $\frac{\partial \mathcal{L}}{\partial \Phi} = 0$, might we have obtained a minimum or saddle point?

B) If $\frac{\partial \mathcal{L}}{\partial \Phi} = 0$ yields a maximum, is it a global maximum?

C) Is $\mathcal{L}$ unimodal?

For the $\mathcal{L}$ functions being considered, (C) can not be answered analytically in the general case. Levin and Reeds (1977) have shown that if $\mu_1, \ldots, \mu_{k-1}$ are known and θ is unknown, then $\mathcal{L}$ has at most one mode. This result suggests that unimodality may be preserved even if the μ_i's are unknown.

Assuming that unimodality cannot be verified, we must address case (B), the arrival at a local maximum. The class of iterative algorithms,

$$\Phi^{(r+1)} = \Phi^{(r)} - M_{(r)}^{-1} \cdot \left(\frac{\partial \mathcal{L}}{\partial \Phi}\right)\Big|_{\Phi = \Phi^{(r)}} \qquad r = 0,1,2,\ldots \tag{3.23}$$

includes both MNR and MEM. Using (3.23) along with a Taylor expansion for $\mathcal{L}(\Phi^{(r+1)})$ at $\Phi^{(r)}$ we obtain

$$\mathcal{L}(\Phi^{(r+1)}) = \mathcal{L}(\Phi^{(r)}) - \left(\frac{\partial \mathcal{L}}{\partial \Phi}\right)^T \cdot (M_{(r)}^{-1}) \cdot \left(\frac{\partial \mathcal{L}}{\partial \Phi}\right)\Big|_{\Phi = \Phi^{(r)}} \tag{3.24}$$

If $M_{(r)}$ is positive (or negative) definite, then the iterative algorithm (3.30) is a descent (or ascent) algorithm. For simplicity we will call this type of algorithm a monotone algorithm. Under a monotone algorithm, the answer to (B) is yes. For the MNR algorithm, this requires that all diagonal elements of $M_{(r)}$ are positive.

It is interesting to note that the choice of reparametrization from $\underline{\alpha}$ to $(\underline{\mu}, \theta)$ makes a difference in the shape of the likelihood and in terms of the behavior of the iterative algorithm. This property reinforces the use of "stable parameters" as justified in a slightly different manner by Ross

(1970). More precisely, we notice that the $\mathscr{L}$ function for both Dirichlet multinomial and Dirichlet negative multinomial has the following form under the $(\alpha_1, \ldots, \alpha_k)$ parametrization

$$\mathscr{L} = c(\underline{x}_t) + \sum_t \sum_i \sum_r \log(\alpha_i + r) - \sum_t \sum_r \log(\Sigma\alpha + r)$$

If we now look at

$$\frac{\partial^2(-\mathscr{L})}{\partial \alpha_i^2} \qquad i = 1,2,\ldots,k$$

we have

$$\frac{\partial^2(-\mathscr{L})}{\partial \alpha_i^2} = \sum_t \sum_i \sum_r \frac{1}{(\alpha_i + r)^2} - \sum_t \sum_r \frac{1}{(\Sigma\alpha + r)^2} \qquad i = 1, \ldots, k$$

From this relation (for all i) it is unclear what the likelihood surface might look like and whether we would obtain a monotone algorithm. Figure 1 illustrates the behavior of the $\mathscr{L}$ function for $k = 2$, $m = 10$, $n = 40$ based upon the sample $x_1, \ldots, x_{10} =$ 23,31,1,1,3,34,17,32,31,8.

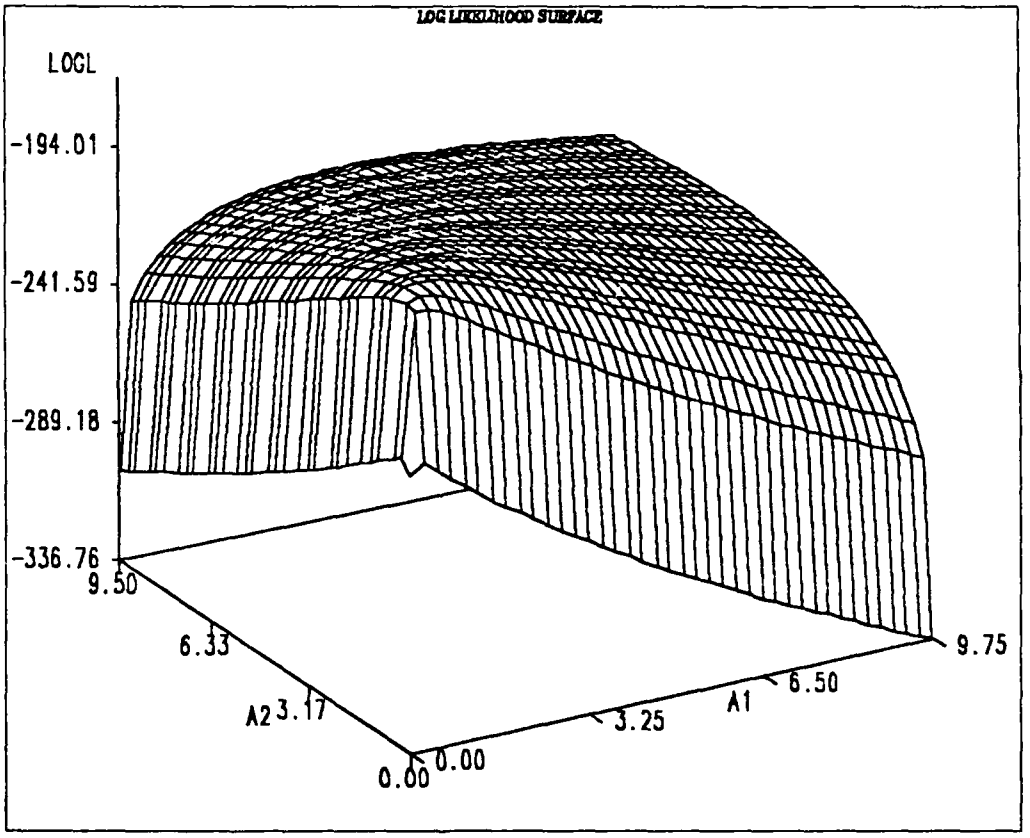

Figure 1. Log Likelihood function under the original parametrization

Under the $(\mu_1, \ldots, \mu_{k-1}, \theta)$ parametrization $\mathscr{L}$ has the form

$$\mathscr{L} = c(\underline{x}_t) + \sum_t \sum_i \sum_r \log(\mu_i + r\theta) - \sum_t \sum_r \log(1 - \sum_{j=1}^{k-1} \mu_j + r\theta) - \sum_t \sum_r \log(1 + r\theta)$$

Now if we examine the second order partial derivatives on the μ_i coordinates

$$\frac{\partial^2(-\mathcal{L})}{\partial\mu_i} = \sum_t \sum_{i} \sum_r \frac{1}{(\mu_i + r\theta)^2} + \sum_t \sum_r \frac{1}{(1 - \sum_{j=1}^{k-1} \mu_j + r\theta)^2} > 0 \quad i = 1, \ldots, k$$

so the log likelihood is concave in all μ directions. Figure 2 illustrates this behavior and we can also see that that $\mathcal{L}$ appears smoother under the (μ, θ) parametrization. To assess whether the iterative algorithm is monotone we need only examine whether $\frac{\partial^2(-\mathcal{L})}{\partial\theta^2} > 0$

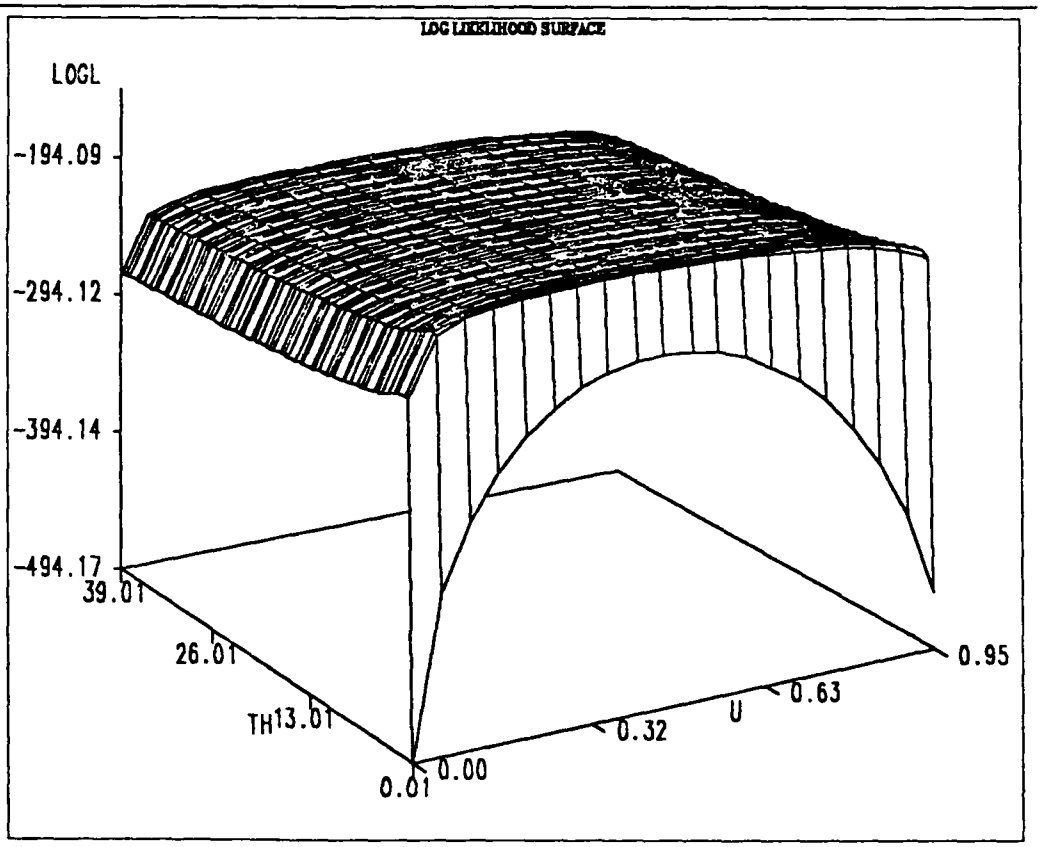

Figure 2. Log Likelihood function under the reparametrization

4. A SIMULATION STUDY

An extensive simulation study was undertaken to compare the method of moments estimators, the MLE's and the jackknifed MLE's (JMLE). Data was simulated using IMSL routines GGAMR and GGUBS. Roughly 5000 replications were used. First, we compare MME's with MLE's. It is expected that the MLE will perform better, and in fact, this is case, often substantially so. For both the DM and DNM case, the parameters to vary are

k: dimension of the parameter space
m: the sample size
n: generalized Bernoulli trial stopping parameter
and α : the Dirichlet parameter vector

We present results for few of the $k = 2,3,6$, $m = 10,40$ and $n = 40,60$ cases in the following tables. (See Leeds (1987) for additional simulation work). For each estimator under each specification we can compute (i) bias, (ii) mean squared error, MSE, (iii) quadratic expected loss using information weighted loss, QEL, (iv) the exact covariance matrix, and (v) the inverse of Fisher's information matrix. BIASM, MSEM, QELM are for the MME; BIAS, MSE, QEL are for the MLE's. Since we know the true $\underline{\alpha}$, we can compute the asymptotic covariance of the MLE's which is $m^{-1} I^{-1}(\underline{\alpha})$ to compare with the exact covariance at the fixed m. This is of interest since we do not know whether the MLE's for these Dirichlet mixed models are asymptotically efficient. Generally, we can not verify the usual regularity conditions (see Lehmann (1980)).

In the last few tables we present simulation results again based on 5000 replications for a few cases ($k = 2$) comparing the JMLE with the MLE.

In summary :

(i) Small sample case m = 10. For both DM and DNM cases, a feasible MME provided an accurate starting solution for the MNR algorithm. However, $\theta \to 0$ or increasing k were more dependent on MEM backup. When MME's did not exist, the starting solution was chosen to be

$$\Phi^{(0)} = (\varepsilon, \varepsilon, \ldots, \varepsilon, \frac{1}{k})$$

This solution performed admirably when the MME did not exist. The asymptotic covariance approximation is poor.

(ii) Large Sample case m = 40. For both DM and DNM compounds, the moments estimate was always feasible and provided a good starting solution for the MNR algorithm. This comes as no surprise since the MME's are consistent. However, if we were to receive a set of data with unequal n_i's , we would not have a MME to start the MNR algorithm. The starting value in (i) should suffice. Here the asymptotic covariance approximation seems more reasonable.

(iii) For the $k = 2$ case Shenton (1950) reports that the efficiency of the MME to MLE's is at least 70% . This result is not contradicted. If we happen to be in this case we do not lose much by using the MME. However, for larger k there appears to be a rather dramatic reduction is efficiency of MME's compared to MLE's.

(iv) JMLE's performed exceptionally well in comparison to the MLE in terms of reducing mean squared error. This result is encouraging for the reason that the evaluation of the JMLE makes the most use of the MLE algorithm to ensure the successful computation of all pseudo values, and thus the JMLE itself. However, it should be mentioned that with frequent use of the MLE algorithm, convergence may be a problem, especially in the small sample case. For instance, if $m = 10$ we compute each pseudo value on nine observations. This is a 10% reduction in the amount of data considered. For the large sample case this would not be so.

α_1	α_2	m	n
1	1	10	40

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
1.513510	1.057329	1.398952	0.991609	8.951	-6.21	.215	.149
1.057329	1.239158	0.991609	1.168164	-6.20	8.951	.149	.215

	BIASM	BIAS	MSEM	MSE
α_1	0.414802	0.443557	1.685580	1.5956954
α_2	0.400124	0.429745	1.399257	1.3528449

MSEM = 3.084837 MSE = 2.948540
QELM = 24.835702 QEL = 24.125809

α_1	α_2	m	n
1	1	40	40

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
0.074439	0.053271	0.068876	0.047896	35.802	-24.8	.054	.037
0.053721	0.073378	0.047896	0.067768	-24.83	35.80	.037	.054

	BIASM	BIAS	MSEM	MSE
α_1	0.060176	0.065845	0.078060	0.0732114
α_2	0.060324	0.066177	0.077017	0.0721478

MSEM = 0.155077 MSE = 0.145359
QELM = 52.383275 QEL = 52.261535

α_1	α_2	m	n
1	5	10	40

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
1.469407	7.120741	1.240462	5.880053	11.04	-1.593	.027	1.244
7.120741	50.092428	5.880053	40.67265	-1.59	11.04	1.244	8.617

	BIASM	BIAS	MSEM	MSE
α_1	0.603724	0.476951	1.834133	1.4679439
α_2	3.684889	2.981112	63.678252	49.5596842

MSEM = 65.512385 MSE = 51.027628
QELM = 28.428756 QEL = 26.017101

α_1	α_2	m	n
1	5	40	40

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
0.122884	0.601518	0.092132	0.441677	44.16	-6.37	.068	0.311
0.601518	4.082731	0.441677	3.162800	-6.374	1.384	.311	2.155

	BIASM	BIAS	MSEM	MSE
α_1	0.114245	0.085229	0.135936	0.0993962
α_2	0.685972	0.547638	4.553290	3.4627071

MSEM = 4.689226 MSE = 3.562103
QELM = 67.364333 QEL = 66.682778

α_1	α_2	α_3	m	n
1	1	1	10	60

Method of Moments Covariance Matrix

0.653 0.428 0.415
0.428 0.714 0.470
0.415 0.470 0.693

Maximum Likelihood Covariance Matrix

0.335 0.188 0.187
0.188 0.334 0.200
0.187 0.200 0.339

Fisher's Information

11.13 -3.789 0.138 .072
11.13 0.138
-3.789 11.13 .072 0.138

Fisher's Inverse

	BIASM	BIAS	MSEM	MSE
α_1	0.474866	0.242310	0.879144	0.3937210
α_2	0.475535	0.240140	0.939910	0.3910328
α_3	0.475880	0.247340	0.919547	0.4004014

MSEM = 2.738601 MSE = 1.186056
QELM = 62.345719 QEL = 45.055172

α_1	α_2	α_3	m	n
1	1	1	40	60

Method of Moments Covariance Matrix

0.056 0.032 0.032
0.032 0.056 0.033
0.032 0.033 0.058

Maximum Likelihood Covariance Matrix

0.042 0.022 0.022
0.022 0.040 0.022
0.022 0.022 0.041

Fisher's Information

44.52 -15.16 .035 .018
44.52 .035
-15.16 44.52 .018 .035

Fisher's Inverse

	BIASM	BIAS	MSEM	MSE
α_1	0.079744	0.048335	0.062908	0.0446581
α_2	0.078848	0.048255	0.061938	0.0421660
α_3	0.078873	0.047966	0.063734	0.0437279

MSEM = 0.188580 MSE = 0.130552
QELM = 95.803146 QEL = 94.523415

α_1	α_2	α_3	m	n
1	3	5	10	60

Method of Moments Covariance Matrix

0.938	2.293	3.884
2.293	10.12	14.89
3.884	14.89	27.13

Maximum Likelihood Covariance Matrix

0.461	1.063	1.815
1.063	4.757	6.833
1.815	6.833	12.75

Fisher's Information

11.5	-1.02	
	2.36	
-1.02	.895	

Fisher's Inverse

.164	.308	.543
.308	1.43	2.00
.543	2.00	4.05

	BIASM	BIAS	MSEM	MSE
α_1	0.622727	0.297014	1.325775	0.5488681
α_2	2.055143	1.002653	14.342457	5.7625477
α_3	3.478274	1.721109	39.231529	15.715182

MSEM = 54.899761 MSE = 22.026598
QELM = 66.288906 QEL = 56.062711

α_1	α_2	α_3	m	n
1	3	5	40	60

Method of Moments Covariance Matrix

0.065	0.134	0.232
0.134	0.620	0.887
0.232	0.887	1.674

Maximum Likelihood Covariance Matrix

0.049	0.098	0.171
1.063	4.757	6.833
1.815	6.833	12.75

Fisher's Information

45.9	-4.12	
	9.42	
-4.12	3.58	

Fisher's Inverse

.041	.077	.136
.077	.359	.501
.136	.501	1.01

	BIASM	BIAS	MSEM	MSE
α_1	0.098757	0.048925	0.074585	0.0510555
α_2	0.327434	0.180411	0.726720	0.4933342
α_3	0.544873	0.299283	1.971414	1.3429237

MSEM = 2.772719 MSE = 1.8873134
QELM = 267.365122 QEL = 265.9265337

α_1	α_2	α_3	α_4	α_5	α_6	m	n
1	1	1	1	1	1	10	60

Fisher's Information

11.89	-1.66	-1.66	-1.66	-1.66	-1.66
-1.66	11.89	-1.66	-1.66	-1.66	-1.66
-1.66	-1.66	11.89	-1.66	-1.66	-1.66
-1.66	-1.66	-1.66	11.89	-1.66	-1 66
-1.66	-1.66	-1.66	-1.66	11.89	-1.66
-1.66	-1.66	-1.66	-1.66	-1.66	11.89

Information Inverse

.107984					.034179
	.107984				
		.107984			
			.107984		
				.107984	
.034179					.107984

Method of Moments Covariance Structure

1.1084	.6072	.5839	.6386	.6566	.6053
.6072	1.0835	.5656	.6144	.6305	.6134
.5839	.5656	1.0680	.6068	.6242	.6028
.6386	.6144	.6068	1.1713	.6543	.6308
.6566	.6305	.6242	.6543	1.1769	.6602
.6053	.6134	.6028	.6308	.6602	1.1105

Maximum Likelihood Covariance Structure

.1865	.0721	.0670	.0718	.0730	.0678
.0721	.1912	.0698	.0726	.0746	.0732
.0670	.0698	.1858	.0715	.0717	.0681
.0718	.0726	.0715	.1880	.0702	.0680
.0730	.0746	.0717	.0702	.1904	.0695
.0678	.0732	.0681	.0680	.0695	.1809

	BIASM	BIAS	MSEM	MSE
α_1	1.078768	0.177409	2.292481	0.2179630
α_2	1.088996	0.182866	2.389933	0.2245902
α_3	1.077059	0.170849	2.248338	0.2150239
α_4	1.082371	0.179794	2.236290	0.2203275
α_5	1.093295	0.175182	2.392863	0.2211166
α_6	1.074797	0.183474	2.285986	0.2146019

MSEM = 13.8728295 MSE = 1.3136232
QELM = 124.469340 QEL = 60.31711222

α_1	α_2	α_3	α_4	α_5	α_6	m	n
1	1	1	1	1	5	10	60

Fisher's Information

```
11.29      -.90778
   11.29
      11.29
         11.29
            11.29
-.90778           .9853
```

Information Inverse

```
.12301 .04016  . . . . .26463
.04016 .12301
.      .   .
.      .   .
.      .   .12301
.26463 .04016       2.23381
```

Method of Moments Covariance Structure

```
2.056
1.222 1.9842
1.210 1.1634 1.945
1.223 1.2689 1.166 2.086
1.213 1.2033 1.167 1.222 1.949
7.610 7.3565 7.232 7.709 7.333 49.24
```

Maximum Likelihood Covariance Structure

```
.2401
.0968 .2347
.0944 .0909 .2260
.0972 .0968 .0877 .2332
.0976 .0980 .0881 .0914 .2281
.6347 .6131 .5841 .6212 .6067 4.8
```

	BIASM	BIAS	MSEM	MSE
α_1	1.394611	0.203828	4.00442	0.2816740
α_2	1.383026	0.205197	3.89698	0.2768250
α_3	1.378675	0.198051	3.84595	0.2652151
α_4	1.380749	0.199832	3.99241	0.2733169
α_5	1.381548	0.200946	3.85799	0.2684726
α_6	7.250609	6.082229	101.81507	6.0822289

MSEM = 121.41281 MSE = 7.4477325
QELM = 169.84660 QEL = 74.8955841

α_1	α_2	α_3	α_4	α_5	α_6	m	n
1	1	1	3	3	3	10	60

Fisher's Information

```
10.92      -.72916
   10.92
      10.92
          2.486
             2.486
-.72916          2.486
```

Information Inverse

```
.1265  .0407  .0407  . . . . . .
.0407  .1265  .0407   .147374
.0407  .0407  .1265
.      .      .845   .534  .534
.  .147374   .534   .845  .534
. . . . . .   .534   .534  .845
```

Method of Moments Covariance Structure

```
2.211
1.352 2.1805
1.428 1.3051 2.285
4.989 4.7342 4.947 19.97
4.718 4.5615 4.684 16.58 18.61
4.924 4.6910 4.815 17.07 16.41 19.39
```

Maximum Likelihood Covariance Structu:

```
.2296
.0931 .2498
.0925 .0932 .2475
.3397 .3551 .3415 1.820
.3215 .3423 .3179 1.216     7
.3368 .3546 .3391 1.272 1.215 1.8
```

	BIASM	BIAS	MSEM	MSE
α_1	1.391406	0.202547	4.14743	0.2706460
α_2	1.387812	0.208158	4.10649	0.2931731
α_3	1.390962	0.205864	4.21970	0.2898637
α_4	4.334136	0.673500	38.75543	2.2733703
α_5	4.269258	0.649777	36.84389	2.1189765
α_6	4.345345	0.682773	38.27363	2.2626982

MSEM = 126.34657 MSE = 7.508727
QELM = 192.06151 QEL = 94.784289

α_1	α_2	m	n
1	3	10	30

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
2.3209	6.7355	1.3750	4.2090	10.71	-2.604	.239	0.601
6.7355	33.2497	4.2090	18.9501	-2.60	1.037	.601	2.474

	BIASM	BIAS	MSEM	MSE
α_1	1.396165	0.485078	4.270152	1.6103188
α_2	4.011426	1.791021	49.341280	22.1579009

MSEM = 53.611432 MSE = 23.768220
QELM = 48.274695 QEL = 29.400234

α_1	α_2	m	n
1	3	40	30

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
0.1839	0.4094	0.0823	0.2195	42.86	-10.42	.060	0.150
0.4094	1.6422	0.2195	0.9222	-10.417	4.149	.150	0.618

	BIASM	BIAS	MSEM	MSE
α_1	0.511964	0.077721	0.446043	0.0883497
α_2	1.242014	0.292713	3.184778	1.0078931

MSEM = 3.63082 MSE = 1.096243
QELM = 79.22263 QEL = 67.379283

α_1	α_2	m	n
1	5	10	30

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
3.0969	15.8208	1.9116	10.1158	10.487	-1.57	0.2899	1.2976
15.8208	115.0248	10.1158	73.1788	-1.572	0.351	1.2976	8.6557

	BIASM	BIAS	MSEM	MSE
α_1	1.239984	0.575918	4.634425	2.2432696
α_2	6.766865	3.659899	160.815227	86.5736542

MSEM = 156.4496 MSE = 88.816923
QELM = 89.2794 QEL = 53.809407

α_1	α_2	m	n
1	5	40	30

Method of Moments Covariance Matrix		Maximum Likelihood Covariance Matrix		Fisher's Information		Fisher's Inverse	
0.2066	0.9496	0.1074	0.5715	41.949	-6.29	0.0725	.32440
0.9496	5.9745	0.5175	3.5226	-6.289	1.405	0.3244	2.1639

	BIASM	BIAS	MSEM	MSE
α_1	0.344309	0.098942	0.3251636	0.1171760
α_2	1.657210	0.597938	8.7208346	3.8801412

MSEM = 9.04599 MSE = 3.9973172
QELM = 69.64925 QEL = 65.9877250

α_1	α_2	α_3	m	n
3	3	3	10	30

Method of Moments Covariance Matrix

11.891	8.646	7.3191
8.646	11.07	7.1634
7.319	7.163	8.9517

Maximum Likelihood Covariance Matrix

3.417	2.861	2.824
2.861	3.411	2.512
2.824	2.812	3.438

Fisher's Information

2.45	-1.02	-1.02
-1.02	2.452	-1.02
-1.02	-1.02	2.622

Fisher's Inverse

.923	.635	.6051
.635	.923	.6051
.605	.605	.8513

	BIASM	BIAS	MSEM	MSE
α_1	3.989816	0.887277	27.810048	4.2047086
α_2	3.927451	0.873922	26.493368	4.1745338
α_3	3.072373	0.870821	18.391223	4.1963659

MSEM = 72.694639 MSE = 12.575608
QELM = 107.797086 QEL = 64.542774

α_1	α_2	α_3	m	n
1	3	5	40	10

Method of Moments Covariance Matrix

0.226	0.523	0.8543
0.523	2.202	3.0777
0.854	3.077	5.2502

Maximum Likelihood Covariance Matrix

0.122	0.295	0.485
0.295	1.167	1.685
0.485	1.685	2.988

Fisher's Information

30.72	-3.19	-3.19
-3.19	6.668	-3.19
-3.19	-3.19	2.906

Fisher's Inverse

.076	.161	.2611
.161	.657	.8982
.261	.898	1.617

	BIASM	BIAS	MSEM	MSE
α_1	0.401361	0.111901	0.3870097	0.1346381
α_2	1.252126	0.367740	3.7695374	1.3020395
α_3	1.820276	0.601499	8.5636105	3.3473450

MSEM = 12.720158 MSE = 4.780227
QELM = 157.781792 QEL = 151.413406

α_1	α_2	α_3	α_4	α_5	α_6	m	n
1	1	1	1	1	5	10	10

Fisher's Information

```
 7.7653        -0.71235
    7.7653
       7.7653
          7.7653
             7.7653
-0.71235             0.81150
```

Information Inverse

```
.20614 .08818 .. .08818  .49059
.08818 .20614    .08818  .49059
   .      .  .      .
   .      .  .      .
.08818 .08818    .20614  .49059
.49059 .49059 ..... 3.38551
```

Method of Moments Covariance Structure

```
127.6
103.0 138.50
 90.3  89.87 103.3
 99.9  99.78  90.7 124.2
 95.4  95.02  87.5  93.8 111.6
436.8 447.59 404.9 446.2 411.5 2054.
```

Maximum Likelihood Covariance Structure

```
2.462
2.049 2.394
2.618 2.556 3.621
1.573 1.524 1.925 1.712
3.642 3.471 4.537 2.254 7.524
17.21 16.64 21.57 12.06 31.3 146
```

	BIASM	BIAS	MSEM	MSE
α_1	5.930699	0.444315	162.82029	2.6592233
α_2	5.771361	0.413110	171.86984	2.5647757
α_3	5.736776	0.436712	136.22782	3.8115527
α_4	5.856497	0.414522	158.52570	1.8840746
α_5	5.856497	0.443810	144.12501	7.7205665
α_6	25.436602	2.407225	2701.14373	151.8331676

MSEM = 3473.71240 MSE = 170.473361
QELM = 2313.20929 QEL = 129.862213

JACKKNIFE RESULTS (DIRICHLET-MULTINOMIAL TABLES k = 2)

α_1	α_2	m	n
1	1	10	40

Biases

	JMLE	MLE
α_1	-0.2265	.4275
α_2	-0.2248	.4161

Mean Squared Error

	JMLE	MLE
α_1	0.3729	1.324
α_2	0.3550	1.153

α_1	α_2	m	n
1	5	10	40

Biases

	JMLE	MLE
α_1	-0.4358	0.5030
α_2	-2.4852	3.1318

Mean Squared Error

	JMLE	MLE
α_1	0.3714	1.743
α_2	10.9462	63.054

α_1	α_2	m	n
0.5	1	10	10

Biases

	JMLE	MLE
α_1	-0.2040	.2425
α_2	-0.4619	.6049

Mean Squared Error

	JMLE	MLE
α_1	0.0800	0.4546
α_2	0.4322	3.1011

α_1	α_2	m	n
0.5	0.5	10	10

Biases

	JMLE	MLE
α_1	-0.1668	0.1963
α_2	-0.1663	0.1953

Mean Squared Error

	JMLE	MLE
α_1	0.0798	0.3839
α_2	0.0801	0.3605

REFERENCES

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Soc.*, **41** ,164-171.

Dalal, S.R., and Hall, W.J. (1983). Approximating priors by mixtures of natural conjugate priors. *JRSSB* , **45** ,278-286.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the E.M. algorithm (with discussion). *J.R.Statist.Soc.B.* , **39** ,1-38.

Efron, B. (1982). The Jackknife, The Bootstrap and Other Resampling Plans. *Siam* , **38**

Everitt, B.S., and Hand, D.J. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

Griffiths, D.A. (1973). Maximum likelihood estimation for the beta-binomial distribution of the total number of cases of a disease. *Biometrics* , **29** , 637-48.

Johnson, N.L., and Kotz, S. (1969). *Discrete Distributions.* Boston : Houghton and Mifflin.

Kalwani, Manohar U. (1980). Maximum Likelihood estimation of the Multinomial-Dirichlet Distribution. *Paper No. 741 (September), Krannert Graduate School of Management, Purdue University* .

Kleinman, J.C. (1973). Proportions with extraneous variance : Single and independent samples. *JASA* , **68** ,46-54.

Leckenby, J.D., and Kishi S. (1982). The Dirichlet Multinomial distribution as a media exposure model. *J.Marketing Research* , **21** ,100-105.

Leeds, S. (1987). Inference for Dirichlet Mixed Distributions. Unpublished Ph.D. dissertation, University of Connecticut.

Lehmann, E.L. (1980). *Theory of Point Estimation.* Wiley, New York.

Levin, B., and Reeds, J. (1977). Compound multinomial likelihood functions are unimodal : Proof of a conjecture of I.J. Good. *Ann. Statist.* , **5** ,79-87.

Mosimann, J.E. (1962). On the compound multinomial distribution, the multivariate beta distribution, and correlations among proprotions. *Biometrika* , **49** , 65-82.

Mosimann, J.E. (1963). On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. *Biometrika* , **50** , 47-54.

Orchard, T., and Woodbury, M.A., (1972). A missing information principle : Theory and Applications. *Proc. Sixth Berkeley Symp. Math Stat and Prob.* , **1** ,697-715.

Pearson K. (1894). Contribution to the Mathematical Theory of Evolution. *Phil. Trans. Roy. Soc. A* , **185** ,71-110.

Quenouille, M. (1949). Approximate tests of correlation in time series. *JRSSB* , **11** ,pp. 18-84.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika* , **43** , pp. 353-360.

Ross, G.J.S. (1970). The efficient use of function minimisation in non-linear maximum likelihood estimation. *Appl. Statist.* , **19** ,205-221.

Rust, R.T., and Leone, R.P. (1984). The mixed-media Dirichlet multinomial distribution: A model for evaluating television-magazine advertising schedules. *J.Marketing Research* , **21** , 89-99.

Schucany W., Gray, H., and Owen (1971). On bias reduction in estimation. *JASA* , **66** , pp. 524-533.

Shenton, L.R. (1950). Maximum likelihood and the efficiency of the method of moments. *Biometrika* , **37** ,111-116.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of a success as variable between the sets of trials. *J.R.Statist.Soc.B.*, **39** , 1-38.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* , **1** , 49-58

Teicher, H. (1961). Identifiability of Mixtures. *Ann. Math. Statist.* , **31** ,55-73.

Teicher, H. (1963). Identifiability of Finite Mixtures. *Ann. Math. Statist.* , **34** ,1265-1269.

Titterington, D.M., Makov U.E., and Smith A.F.M. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley and Sons Ltd.

Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* , **31** , 949-952.

Wu, C.F. (1983). On the convergence properties of the EM algorithm *Ann. Statist* , **11** , 95-103.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 428	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Estimation For Dirichlet Mixed Models		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Steve Leeds and Alan E. Gelfand		8. CONTRACT OR GRANT NUMBER(s) N00014-89-J-1627
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE May 19, 1990 13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Dirichlet mixed models, maximum likelihood estimation, moment estimators, comparison of estimators.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE FOLLOWING PAGE.

TECHNICAL REPORT NO. 428

20. *ABSTRACT

Dirichlet mixed models find wide application. Estimation is usually achieved through the method of moments. Here we present an iterative hybrid algorithm for obtaining the maximum likelihood estimate employing both modified Newton-Raphson and E-M methods. This successful MLE algorithm enables calculation of a jackknife MLE. Simulation comparison of the three estimates is provided. The MLE substantially improves upon the moments estimator particularly with increasing dimension. The jackknife MLE in turn offers dramatic improvement over the MLE.